Does the visual channel improve the perception of consonants produced by speakers of French with Down syndrome?

Alexandre Hennequin[1]

Amélie Rochet-Capellan[1]

Silvain Gerber[1]

Marion Dohen[1]

[1] Univ. Grenoble Alpes, CNRS, GIPSA-lab, F-38000 Grenoble, France

Correspondance concerning this article should be addressed to Marion Dohen, Department of Speech and Cognition, Laboratoire Gipsa-lab, Domaine Universitaire, BP 46, 38402 Saint Martin d'Hères cedex, France. E-mail: marion.dohen@gipsa-lab.grenoble-inp.fr

Abstract

Purpose: This work evaluates whether seeing the speaker's face could improve the speech intelligibility of adults with Down syndrome (DS). This is not straightforward since DS induces a number of anatomical and motor anomalies affecting the orofacial zone.

Method: A speech-in-noise perception test was used to evaluate the intelligibility of 16 consonants (C) produced in a VCV context (V = /a/) by 4 speakers with DS and 4 control speakers. 48 naïve participants were asked to identify the stimuli in 3 modalities: Auditory (A), Visual (V) and Auditory-visual (AV). The probability of correct responses (PCR) was analyzed as well as AV gain, confusions and transmitted information as a function of modality and phonetic features.

Results: PCR follows the trend AV > A > V, with smaller values for the DS than the control speakers in A and AV but not in V. This trend depended on the consonant: the visual information particularly improved the transmission of place of articulation and to a lesser extent of manner, while voicing remained specifically altered in DS.

Conclusions: The results suggest that the visual information is intact in the speech of people with DS and improves the perception of some phonetic features in consonants in a similar way as for control speakers. This result has implications for further studies, rehabilitation protocols and specific training of caregivers.

Introduction

Managing to produce intelligible speech sounds is a challenge for people with Down-Syndrome (DS). As a result, parents, speech therapists and researchers in speech sciences try to provide them with appropriate help, primarily oriented toward diagnostic and improvement of acoustic intelligibility (Kent & Vorperian, 2013; Kumin, 2012; Meyer, Theodoros, & Hickson, 2016). It is however well known that, in face-to-face communication, people do not only use acoustic information to process speech but also visual information (e.g. lip reading). This information is particularly useful when the acoustic signal is degraded, as is the case in noisy environments (e.g. Schwartz, Berthommier, & Savariaux, 2004), but also for speech produced by a non-native speaker (Reisberg, McLean, & Goldfield, 1987) or for dysarthric speech (Borrie, 2015; Hustad, Dardis, & McCourt, 2007). Visual information could therefore also improve the intelligibility of speakers with DS.

DS however induces craniofacial, occlusal and dental anomalies as well as weak and poorly differentiated intra-oral and facial muscles (Arumugam et al., 2015; Kent & Vorperian, 2013). These specificities could affect the visual and audio information conveyed during speech production in DS. Is the visual information preserved in speech produced by people with DS? Are some speech sounds better perceived when listeners can see the speakers' face? In particular, what is the contribution of the visual channel to the perception of consonants and the transmission of phonetic features? In this paper, we address these issues by analyzing the perception, by non-familiarized participants, of vowel-consonant-vowel sequences produced by young adults with DS vs. control speakers.

What is Down Syndrome?

Down Syndrome (DS) is a common genetic condition related to the presence of an extra chromosome 21. It is the best-known genetic origin of intellectual deficiency (Katz & Lazcano-Ponce, 2008). DS is present worldwide but its live births prevalence varies depending on the country, mainly in relation to maternal age, health care facilities and fetal termination politics (cf. Loane et al., 2013; Parker et al., 2010). As an illustration, DS concerned ~570 newborns in France in 2012 (HAS, 2015) and ~5657 newborns per year in the United States in 2004-2006 (Parker et al., 2010). The life expectancy of people with DS has increased from 12 years in 1940 to about 60 years nowadays (Bittles et al., 2007). Providing adapted medical care, educational support and promoting the social integration of these persons are worldwide challenges. Improving their communication is part of it.

Intelligibility of speech produced by people with DS

Speech intelligibility is frequently reported as impaired for speech produced by people with DS with crucial consequences on their social participation and integration. Parental surveys revealed that, on a 10-point scale (1 corresponding to 'completely unintelligible' and 10 to 'completely intelligible'), intelligibility is rated on average between 4 and 5 (Kumin (2006): n=1,620, age: 1 to 21, mean age: 8.2, USA; Toğram (2015): n=319, age: 1 to 19, mean age: 5.3, Turkey). Only a minority of parents evaluated their child as being completely intelligible (Kumin: 1.5% of the parents; Toğram: 6%). More parents reported systematic or frequent difficulties for consonant (Kumin: 64.7%; Toğram: 45.5%) and to a lesser extent for vowels (Kumin: 42.4%; Toğram: 33.8%) production. In these surveys, parents evaluated the intelligibility of their child based on everyday experience and communication in natural settings. Visual correlates of speech are thus implicitly integrated in these evaluations even if they were not specifically evaluated.

Kent and Vorperian (2013) reviewed the clinical and experimental studies on speech production in DS from 1950 to 2012. They reported that most of the examined studies on speech intelligibility were based on transcriptions of audiotaped speech (narrative, conversational, picture naming…), intelligibility being quantified as the proportion of complete and intelligible utterances. The percentage of correct consonants (PCC), calculated based on transcriptions by speech therapists, was also a frequent indicator since "it has been found to be correlated with speech intelligibility" and "is a good index of speech disorder severity" (Barnes et al., 2009). Based on these measures, acoustic intelligibility was found to be reduced in children and/or adolescents with DS when compared to typical speakers matched in non-verbal mental age (Rupela, Velleman, & Andrianopoulos, 2016; see Kent & Vorperian, 2013 for a review of studies before 2012). Intelligibility of children with DS also appears to improve with chronological age (Chapman & Hesketh, 2001; Rosin et al., 1988). Surprisingly, as underlined by Kent and Vorperian (2013), few studies used methods from speech production and perception research to investigate the phonetic intelligibility of people with DS.

Bunton, Leddy and Miller (2007) audiotaped 5 male adult speakers with DS while they were producing lists of words chosen to evaluate 19 minimal-paired phonological contrasts (single word production). The productions were then transcribed by 5 experts and used as stimuli in a multi-choice perception test involving 10 naïve participants. The two groups evaluated overall intelligibility consistently showing high variability between speakers with DS. A detailed analysis suggested that the largest proportions of errors were observed for initial and final clusters, which were often misperceived as singletons. The proportion of errors was also relatively high for pairs contrasting in place of articulation for both stop and fricative manners and for vowels in the front-back, high-low and long-short dimensions. In a following X-ray

study, Bunton and Leddy (2011) analyzed tongue movements during vowel production by 2 speakers with DS. They found a reduced F1/F2 acoustic vowel space in speakers with DS compared to controls as well as a reduction of the articulatory space. A reduced F1/F2 space was also observed for children by Moura et al. (2008).

Based on transcriptions by trained listeners, Timmins et al. (2011) found that /t/ was produced correctly in average in 71.5% of the trials by children with DS (n=26; mean age: ~13) but in 100% of the trials when produced by typically developing children matched in cognitive age. Similarly, Timmins et al. (2009) reported that /□/ was produced correctly in 46% of the trials by children with DS (n=20, mean age ~13), but in more than 90% of the trials in a control group. Children with DS were reported to produce a non-sibilant fricative instead of /□/ but also, to a lesser extent, a nasal, a plosive or a liquid. Liquid and nasal simplifications were also outlined in children with DS in other studies (Crosley & Downling, 1989; Sommers, Patterson, & Wildgen, 1988).

Most of the studies on speech sound disorders in DS (for a more complete analysis, see Kent & Vorperian, 2013), focused on transcriptions by specialists and, more rarely, on acoustical or articulatory analyses or perceptual evaluations by non-specialists. In everyday life, the social integration of people with DS depends on their ability to be understood by non-familiarized listeners. Moreover, to our knowledge, there is no published work systematically analyzing the perception of consonants produced by adult speakers with DS and/or evaluating the contribution of visual information to this perception.

Causes of intelligibility reduction in DS

Speech impairment in people with DS can be linked to various well known types of difficulties induced by the chromosomal aberration including breathing limits, hearing loss, malformations of

speech articulators related to craniofacial anomalies and neuromuscular issues (Kent & Vorperian, 2013). As an illustration, the size of the oral cavity was reported to be smaller in people with DS than in typical individuals (Borghi, 1990), in relation to an underdevelopment of midface bones. By contrast, pharyngeal length and volume (Xue, Kaine, & Ng, 2010) and tongue size (Guimaraes et al., 2008; Macho et al., 2014) were found to be average. Put together, these factors result in an atypical resonance cavity, a well-known relative macroglossia and occlusal/dental anomalies. Movements are also specifically impaired in people with DS. Hypotonia, low muscle tone, is a commonly reported feature that seems to affect all muscles, including facial and intra-oral ones (e.g. Connaghan & Moore, 2014; Latash, Wood, & Ulrich, 2008). All these anomalies contribute to a disorder in speech sound articulation.

Disorders in articulation as well as in prosody, fluency and voice are observed to various degrees in people with DS and all contribute to speech intelligibility reduction (Bunton et al., 2007). A major point is that this intelligibility reduction is not only due to intellectual deficiency but is structural as well (Cleland et al., 2010): receptive speech skills are usually better than expressive ones in people with DS. In a recent paper, Rupela et al. (2016) suggest that the motor disorder of speech production in children with DS is a complex and variable combination of symptoms of childhood apraxia as well as childhood dysarthria and "Motor Speech Disorder–Not Otherwise Specified".

Could visual information help perceive speakers with DS?

Definitions of intelligibility usually include the listener. Hence it could be "broadly defined as the accuracy with which a listener is able to decode the acoustic signal of a speaker" (Hustad & Cahill, 2003). But, as underlined by De Gelder & Bertelson (2003), in everyday life, perceivers always combine different sensory inputs to make perceptual judgments. This is all the more true

for speech: it is not only heard, it is also seen. The role of visual information in speech perception has been well established: when we look at a speaker while listening to her, what we perceive is actually an integration or binding of visual and auditory information (Massaro, 1987; reviews: Campbell, 2008; Peelle & Sommers, 2015). Not only does seeing the speaker help, for example, identify place of articulation for consonants (Summerfield, 1987), it also provides temporal information on when crucial acoustic cues may occur focusing the listener's attention on these cues (Schwartz et al., 2004) and helping auditory stream segregation (Carlyon, Cusack, Foxton, & Robertson, 2001) and speech detection in noise (Grant & Seitz, 2000).Visual information is particularly relevant in noisy environments, when the quality of the acoustic signal is reduced (Sumby & Pollack, 1954; review: Peelle & Sommers, 2015). Such a paradigm is used very frequently in audiovisual speech perception research in order to put forward visual enhancement avoiding a ceiling effect in the auditory alone modality (e.g., Bernstein, Auer, & Takayanagi, 2004; Sumby & Pollack, 1954). Hence, typical listeners are able to extract featural information from seeing the movements of the speaker's mouth. Some phonetic features have been shown to be more prominent in the auditory channel and others in the visual one. Summerfield (1987) reported that voicing is the most robust feature in the auditory channel, while place of articulation is the most robust in the visual channel. Miller and Nicely (1955) analyzed confusions between 16 English consonants perceived by 5 participants. The auditory signals were degraded with frequency distortion and random masking noise. The authors provide confusion matrices for 5 signal-to-noise ratios (-18, -12, -6, 0, +6 dB) and find that voicing and nasality are quite robust to noise unlike place. Phatak, Lovitt and Allen (2008) also analyzed confusions between English consonants perceived by 24 participants for 5 signal-to-noise ratios (-12, -6, 0, 6, 12 dB; white noise). They found confusion matrices very close to those of Miller and Nicely (1955).

A few studies have investigated the contribution of visual information to the perception of speech in speakers with dysarthria. Keintz, Bunton and Hoit (2007) had 10 experts and 10 inexperienced listeners transcribe sentences produced by eight speakers with Parkinson's disease in auditory (A) and auditory-visual (AV) conditions. Results showed a better intelligibility in AV than in A but only for the less intelligible speakers. Similar observations were made by Hustad and Cahill (2003). Hustad et al. (2007), and more recently Borrie (2015), however found improvement in AV compared with A only for moderate dysarthria. Results concerning speakers with severe dysarthria are inconsistent. Acknowledging the discrepancy of the results in the dysarthric population and the specific anomalies observed in DS and discussed above, it is impossible to predict from the latter results what will be observed in the specific case of DS. Also note that none of the studies described above involved a V only condition making it impossible to assess the quality of the visual information in dysarthric speech. Moreover, they did not provide a specific characterization of the contribution of the visual modality as a function of phonetic feature and did not make direct comparisons with typical speakers. It is also possible that listeners poorly use the visual channel for less severe speech impairments in un-noisy laboratory conditions. This does not mean they do not in everyday life, as speech is often perceived in noisy conditions, and as the effect of this ecological noise might be greater for impaired than typical speech.

The current study analyzes the potential contribution of visual information in the perception of consonants produced by adults with DS by naïve participants using a classic speech-in-noise perception paradigm. The study reported hereafter was designed to address the following questions: (1) If people with DS are less intelligible than control speakers in the auditory modality, is this also the case in the visual modality? (2) Does visual information, when

combined to auditory information, improve the perception of consonants produced by speakers with DS? (3) What are the most frequent errors made in the identification of DS speech in the Audio (A), Audio-Visual (AV) and Visual (V) modalities? How are phonetic features transmitted in DS speech as a function of modality and compared with typical speakers?

Methods

Recording and design of the stimuli for the perception test

Speakers

The speakers, all native speakers of French, were four young adults with DS (DS – two females) and four control speakers (Ctr) matching those with DS in age (±5 years) and gender. Speakers with DS were involved in the study in collaboration with a local association of families (ARIST – www.arist.asso.fr). Control speakers were students recruited via advertisements at Grenoble Alpes University, France. They did not report any history of speech pathology or impairment, nor facial surgery. Table 1 summarizes the main characteristics of the speakers involved in the study. Further information about the speakers with DS is available in supplementary material (S1) and shows that speakers with DS covered a broad range of intelligibility levels.

[Insert Table 1]

All the speakers gave their informed written consent to participate in the study and to be video-recorded, with restricted conditions of use of their videos. For the speakers with DS, both the person and her parent(s) signed the consent and image right forms. The purpose and conditions of the study were orally explained to the person with DS by the experimenter during a video-recorded interview in order for her to give her informed agreement to participate. All the speakers received a 15€ gift card for their participation.

Speech sequences

The speech sequences were sixteen Vowel-Consonant-Vowel (VCV) sequences in which V was always /a/ and C one consonant among /b, d, g, v, z, □, p, t, k, f, s, □, m, n, □, l/. Table 2 summarizes the articulatory features of each consonant. Non-sense sequences were used (as in e.g. Grant, Tufts, & Greenberg, 2007) in order to test pure phonetic intelligibility ruling out semantic and lexical influences.

[Insert Table 2]

Recording procedure

The speakers were recorded in a soundproof room. They wore a head mounted microphone (Sennheiser HSP4) and sat in a chair in front of a loudspeaker and a HD digital camera (Panasonic HC-X920). The field of view of the camera was adjusted from above the head to shoulder level. Audio was sampled at 44100 Hz (Focusrite Scarlett 6i6 soundcard). The speakers heard the VCV sequences, uttered by a different speaker, through the loudspeaker and were instructed to repeat what they had heard. Repetition was chosen, rather than reading, because some speakers with DS were not able to read. Each VCV sequence was produced three times in random order. The audio prompts were recorded from three different female speakers. The three repetitions therefore resulted from repetition after three different speakers. When the speaker failed to produce the right target, the audio prompt was played again until the two experimenters judged that the speaker had uttered her best production of the intended target. This procedure was chosen to reduce perceptual errors. The clearest production was chosen as the VCV exemplar for the perception test, as a tradeoff between auditory and visual quality, and based on the agreement of three of the authors.

All the acoustic stimuli were normalized at 70 dB using Praat (Boersma, 2001). A "cocktail party" noise (BDBRUIT database: Zeiliger et al., 1994) was then mixed with the audio stream at a signal to noise ratio (SNR) of -4 dB. Noise was added in order to avoid a ceiling effect, especially for the control speakers. "Cocktail party" noise was used (rather than white noise for example) for the sake of naturalness (Alm, Behne, & Wang, 2009). The resulting sound files were mixed with the corresponding video files at a 960x540 pixel resolution using FFmpeg (https://www.ffmpeg.org/) to create the Audio-Visual (AV) version of the stimulus. The Audio only (A) version was obtained by replacing the video stream with a static picture of a loudspeaker and the Visual only (V) version by turning the audio stream off. This resulted in a total of 48 stimuli for each of the 8 speakers (3 modalities x 16 VCVs).

Participants in the perception study

48 typical native speakers of French participated in the perception study (24 females and 24 males – age: mean = 24.9, standard error = 3.5). All of them reported normal or corrected to normal vision, no auditory problems and no speech disorder or phonological issues. Before the experiment, each participant underwent a bilateral hearing test consisting of pure tone hearing at 30 dB for 500 Hz, 1 kHz, 2 kHz and 4 kHz. This test confirmed that all participants had normal hearing. They all had little or no experience with people with DS and received a 15€ gift card for their participation.

Procedure

In total, 384 stimuli had to be evaluated (48 stimuli x 8 speakers). In order for the duration of the perception test to be reasonable (~45 min), participants were randomly assigned to two separate subtests each consisting of the stimuli of four speakers (2 with DS and 2 Controls, 192 stimuli).

Participants were seated in a quiet room, approximately 60 cm from a 24" screen (Dell S2415H) and wore a headset with headphones and a microphone (Audio Technica BPHS1).

The perception test was programmed using the Psychophysics Toolbox (Brainard, 1997). It was divided into three blocks, one for each modality (A, V and AV), consisting of 64 stimuli each (16 VCVs x 4 speakers). The 6 possible presentation orders of the blocks were balanced across participants and stimulus order within each block was randomized. The organization of one trial is illustrated in Figure 1. An empty gray square first appeared for 1 sec. The stimulus was then played twice in a row, with a pause of 500 ms (black screen) between presentations. Participants gave their response orally when a green screen appeared, after a 1.5 sec pause (red screen). They then hit a key on the keyboard to move to the next trial. Participants' responses were recorded using the microphone. Oral responses were chosen rather than written transcriptions to avoid spelling ambiguities.

[Insert Figure 1]

Before the test phase, participants were trained to the procedure using noiseless stimuli different from those of the experiment. Two stimuli per modality were presented with the same procedure as that of the test. Participants were then informed that the stimuli in the test would be played with a background noise and were familiarized with a sample of this noise.

Instructions

Participants were informed they would hear and/or see an audio or video or audio-video stimulus twice. They were instructed to repeat what they had perceived after the second stimulus, when the green screen appeared. They were told that the stimuli were meaningless speech sequences. No further information, such as the structure of the sequences, was provided.

Transcription of the participants' responses

All the responses were phonetically transcribed and each phoneme was then assigned to one of the following five items:

BeforeV1 – V1 – C – V2 – AfterV2

C could be either a single consonant or a cluster; V1 and V2 a vowel; beforeV1 and afterV2, anything perceived before V1 or after V2. Each item could also be empty. Table 3 provides examples of transcriptions. C was then classified into one of 17 categories: one of the 16 consonants or 'Other' (e.g. cluster, no consonant perceived, no response provided, ambiguous response…). When it was impossible to transcribe one or several of the 5 items, it was annotated as '?'. The first author transcribed all the responses. The last author independently transcribed half the responses. The agreement score between these two annotations was of 97.6%. Another person (speech therapy student) performed independent transcription of the other half of the responses with an agreement score of 96.8%. All transcriptions were performed blindly from stimulus and experimental condition (the transcribers did not know what the initial stimulus was nor the modality it had been presented in). A third person was then asked to choose between the two transcriptions for all disagreements. We kept this choice for the subsequent analyses. When this person did not agree with any of the two annotations (only 5% of the cases), the item was transcribed as '?'. Note that '?' transcriptions correspond to only 1.8% of all transcriptions.

[Insert Table 3]

Data analyses, statistics and hypotheses

All the analyses were run using the R software (version 3.4.2, R Development Core Team, 2008). Statistical tests were considered significant for $p < .05$. The main factors included in the analyses

were: *Modality* (Audio-visual (*AV*) vs. Audio (*A*) vs. Visual (*V*)); speaker group (*Speaker_group*, Down syndrome (*DS*) vs. Control (*Ctr*)); stimulus presented (*Stimulus*, the 16 VCV sequences produced by the speakers); order of presentation (*Pres_order*, *AV/A/V* vs. *AV/V/A* vs. *A/AV/V* vs. *A/V/AV* vs. *V/AV/A* vs. *V/A/AV*); *Speaker (DS1* to *DS4*, and *Ctr1* to *Ctr4*); and *Participant* (48 levels). Consonants were also grouped along three phonetic features for a subpart of the analyses (cf. Table 2): *Voicing* (voiced vs. unvoiced); *Place* of articulation (labial vs. coronal vs. dorsal); *Manner* of articulation (plosive vs. fricative vs. nasal vs. other).

Analysis 1: Probability of correct identification of VCV sequence

We first analyzed the probability of correct responses (*Prob_correct_VCV*) as a function of *Modality* and *Speaker_group* to provide a global picture of VCV intelligibility, independently from error type. The analysis was done regardless of the presentation order of modalities since it was counterbalanced across participants but *Pres_order* effects are available in supplementary material S3. Based on previous work, an *AV* > *A* > V trend was expected for the *Ctr* group, as well as a *Ctr* > *DS* trend in *A*. If the visual information also plays a role in the perception of DS speech, an *AV* > *A* trend should also be observed for *DS* speakers. A core question was then: does the visual information benefit as much for *DS* than for *Ctr*?

The statistical analysis used was a logistic regression (in R: function *glmer* of the package *lme4*, version 1.1.14) since response correctness is a binary variable (correct: response = stimulus – incorrect: response ≠ stimulus). *Modality*, *Speaker_group*, *Stimulus* and their interactions were included as fixed effects and *Speaker* and *Participant* as random effects including random slopes on the effect of *Modality, Speaker_group* and their interaction. The predictive quality of the model was checked by computing the area under the receiver-operating characteristic curve (ROC AUC) from the model, with values greater than 0.7 being considered as fair. Multiple

comparisons were run on the model (using *glht* function of package *multcomp,* Hothorn et al., 2008).

Paired t-tests were used to ensure that the probabilities of correct responses were greater than 1/16 (chance) for all levels of *Modality* and *Speaker_group*. The corresponding Bonferroni correction was then applied to all p-values (multiplication by the number of comparisons, i.e. 6).

Analysis 2: AV gain

A second analysis was performed to examine the effect of *Speaker_group* on AV gain relative to performance in *A*. *AV Gain* was calculated for each participant as follows (Sommers, Tye-Murray, & Spehar, 2005):

$$AV\ Gain = \frac{AV - A}{100 - A}$$

where *AV* and *A* are the participant's scores in the respective modalities. This method was used to withdraw the impact of the participant's performance in *A* especially since we expect it to be *Speaker_group* dependent. *AV Gain* provides a quantification of visual enhancement relative to *A* only perception (Sommers, Tye-murray, & Spehar, 2005).

Analysis 3: Probability of correct identification of C

We also assessed whether the effect of *Modality* and *Speaker_group* depended on the *Stimulus* (16 levels). To do so, we considered the probability of correct identification of the consonant (C), even if V1 and/or V2 were incorrect and/or something was added before V1 and/or after V2 (cf. Table 3). The same statistical analysis as that described for Analysis 1 was used, the only change being the observed variable (probability of correct identification of the consonant instead of *Prob_correct_VCV*). For space reason and overlapped conclusions with confusion matrices (see below), this analysis is provided in supplementary material (S4).

Analysis 4: Confusion matrices

In order to better understand perceptual errors on consonants as a function of *Speaker_group* and *Modality*, confusion matrices (M) were computed for each *Modality*Speaker_group* condition. They are 16x16 matrices in which rows correspond to the stimuli and columns to the responses. $m_{i,j}$ corresponds to the total number of responses j provided for stimulus i, all participants taken together. Note that we considered the number of observations regardless of the participant due to the small number of repetitions (n=2) for each *Stimulus*Modality*Participant* condition.

Analysis 5: Transmitted information (entropy) of phonetic features

The last part of the analyses was dedicated to the "quality" of transmission of the three phonetic features (*Place*, *Manner* and *Voicing*) as a function of *Modality* and *Speaker_group*. The amount of transmitted information was computed for each phonetic feature in each *Modality* and *Speaker_group*. The aim of this analysis is to examine how well a specific feature is transmitted from stimulus to response. Percentage of transmitted information (*I*) was calculated using entropy with the same method as described in Robert-Ribes, Schwartz, Lallouache, & Escudier (1998), with:

$$I = 100\frac{H(s,r)}{H(s)}$$

where $H(s,r)$ is the information shared between stimulus (s) and response (r) and $H(s)$ is the information in the stimulus. The computation is detailed in supplementary material (S2).

Resulting I ranges from 0% (no information transmitted at all from stimulus to response) to 100% (information systematically well transmitted). *I_Place*, *I_Voicing* and *I_Manner* will further refer to the transmitted information for each phonetic feature.

Based on previous work (Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998; Summerfield, 1987) we expected: for the *Ctr* speakers: (1) Greater *I_Place* in *AV* than in *A*; (2) Equivalent *I_Voicing* and *I_Manner* in *AV* and *A*. The remaining questions were then: would similar trends be observed for the speakers with *DS*? What are the most altered features in *DS* speech and in which modality?

These questions were assessed using a beta regression model (function *glmmadmb* of the package *glmmADMB* 2016.0.8.3.3, Fournier et al., 2012).The complete model was used to perform the multiple comparisons, and its predictive quality was assessed in the same way as the model used in Analysis 1. *Modality*, *Speaker_group*, *Feature* and their interactions were included as fixed effects and *Participant* as random effects including random slopes on the effect of *Modality* and *Speaker_group*. Note that *I* values were transformed to be in]0; 1[to fit the requirement of the beta regression (Smithson & Verkuilen, 2006).

Results

Distribution of response types

57.4% of all the responses include at least one error (see Table 4) with more than 93.2% of these responses involving at least an error on the consonant. There are more perception errors for the speech produced by *DS* than *Ctr* speakers. We then analyzed the probability of correct response as a function of experimental condition.

[Insert Table 4]

Probability of correct VCV responses (Prob_correct_VCV – Analysis 1)

The aim of this analysis was to evaluate how accurately VCV sequences were perceived as a function of *Speaker_group*, *Modality* and *Stimulus* (Figure 2). *Prob_correct_VCV* is significantly

above chance in all *Speaker_group* * *Modality* conditions (for the six corrected t-tests: $t(47) > 7.76$, $p < .001$). The area under the ROC curve computed from the full model (logistic regression) is 0.79 (fair predictive level). Multiple comparisons were then run based on this model to analyze the effects of *Modality* as a function of *Speaker_group* and the reverse (see Tables 5 & 6). In summary, the following trends can be extracted for *Prob_correct_VCV*:

- $AV > A > V$ for *Ctr* speakers ($p < .001$ for all comparisons) and $AV > A \sim V$ for *DS* speakers ($p < .001$ except for *A*-*V*: $p = .14$);
- Between group comparisons show significantly better performance for *Ctr* than for *DS* speakers in *A* ($p < .001$) and *AV* ($p = .01$) but equivalent performances in *V* ($p = .97$).

[Insert Figure 2]

[Insert Table 5]

[Insert Table 6]

AV gain (Analysis 2)

The *AV gain* relative to performance in *A* is not significantly different for *Ctr* (mean = 0.39, sd = 0.23) and *DS* (mean = 0.38, sd = 0.17) speakers (paired t-test: $t(47) = 0.313$, $p > 0.7$). Visual enhancement thus appears to be equivalent in both speaker groups (see Figure 3).

[Insert Figure 3]

Probability of correct identification of C (Analysis 3)

We then further analyzed the probability of correctly identifying the consonant regardless of other potential errors. The analysis of the effects of *Modality* and *Speaker_group* for each consonant (cf. supplementary material S4) suggests that the effects depend on consonantal features. In general, labial consonants follow an $AV > A \sim V$ trend while coronals rather follow an

AV ~ *A* > *V* trend and plosive dorsals an *AV* > *A* > *V* trend. Main differences between groups (*DS* < *Ctr*) are observed:

- in the *A* modality for /d/, /□/, /l/, /g/ and marginally significant for /s/ (amplitude of differences: 0.32 < *Ctr-DS* < 0.43);

- in the *AV* modality only for /d/, /z/, /l/ and marginally significant for /□/ (0.31 < *Ctr-DS* < 0.41).

In *V*, absolute differences between groups are always smaller than 0.18 (never significant) regardless of the consonant.

Confusion matrices (Analysis 4)

The aim of this analysis was to examine into more details the types of errors made on consonants as a function of *Modality* and *Speaker_group*. Confusions between consonants correspond to ~98% of the errors on consonants for both groups and are detailed in the confusion matrices displayed in Figure 4. The following trends can be extracted from the analysis:

- *Voicing confusions* follow a trend *AV* ~ *A* < *V* for both *Ctr* and *DS* speakers. They are more frequent for *DS* than *Ctr* speakers in *A* (*DS*: 22%, *Ctr*: 9.6%) and *AV* (*DS*: 20.7%, *Ctr:* 7.4%) but not in *V* (~40% for both groups). Confusions are observed in both directions in *V*: voiced responses for unvoiced stimuli and the reverse. In *AV* and *A*, the tendency is to identify voiced consonants as unvoiced rather than the reverse, cf. /b/ (resp. /d/ and /g/) identified as /p/ (resp. /t/ and /k/).
- *Manner confusions* follow a trend *AV* < *A* < *V* for both speaker groups with no between group differences (*AV*: 16% (*DS*), 11.7% (*Ctr*) – *A*: 28.8%, 24.3% – *V*: 36.8%, 34.2%). These confusions particularly concern nasal consonants (/m/-/n/) in all modalities.

- *Place confusions* follow the trend $AV < V \sim A$ for *Ctr* speakers and $AV < V < A$ for *DS* speakers. They are relatively rare in *AV* for both speaker groups (*DS*: 9.2%, *Ctr*: 5.2%) and comparable between groups in *V* (*DS*: 17.7%, *Ctr*: 19.5%). The main between group difference is observed in *A* with more confusions for *DS* (28.3%) than *Ctr* (20.1%) speakers. In *V*, *Place* confusions are far less frequent than in *A* for *DS* speakers whereas they are relatively as frequent in both modalities for *Ctr* speakers.
- *"Other" responses* are also frequently provided for some consonants (cf. /□/) for both speaker groups in all modalities, the tendency being strongest in *V*.

[Insert Figure 4]

Feature information transmission (Entropy, Analysis 5)

We analyzed the information transmitted for each phonetic feature (see Figure 5). The area under the ROC curve computed from the full model (beta regression) is 0.82 (good predictive level).

I_Voicing – For *Ctr* speakers, *I_Voicing* follows a trend $AV > A > V$, with *AV-V* and *A-V* greater than 55% ($p < .01$), and *AV-A* ~ 8% ($p = .03$). For *DS* speakers, the trend is the same but the only significant difference is between *AV* and *V* (*AV-V* ~ 29%, $p < .01$). *I_Voicing* is also significantly greater for *Ctr* than *DS* speakers in *A* and *AV* ($p < .04$ for both comparisons), but not in *V* ($p = 1$).

I_Manner – For both speaker groups, *I_Manner* follows the trend $AV > A \sim V$: *Ctr*: *AV-A*: 23% ($p < .01$), *AV-V*: 37% ($p < .01$), *A-V*: 15% ($p = .06$); *DS*: *AV-A: 25%* ($p < .01$), *AV-V*: 33% ($p < .01$), *A-V*: 8% ($p = .99$). *I_Manner* is comparable for both speaker groups in *A* and *V* ($p > .6$). The difference in *AV* is marginally significant (*Ctr-DS*: 8%, $p = .09$).

I_Place – For both speakers groups, *I_Place* follows the trend $AV > V \sim A$ for *Ctr* speakers: *AV-V:* 33%*, AV-A: 35%* ($p < .01$ in both cases), *V-A*: 2.2% ($p = 1$); but less clearly for *DS* speakers:

AV-V: 19% (p = .21)*, AV-A*: 44% (p < .01), *V-A*: 25% (p = .6). Differences between speaker groups are not significant (*A*: p = .81 – *V*: p = .99 – *AV*: p = .48). Note that, once again, it appears that whereas in *V*, transmission of place information is equivalent between groups, it is far less efficient in *A* for *DS* than *Ctr* speakers.

[Insert Figure 5]

Discussion

The aim of this study was to characterize the quality of the visual information in speech produced by people with DS as well as its contribution to general intelligibility. In particular, it investigated the role of visual information in the transmission of consonantal phonetic features by speakers with DS as compared with typical speakers matched in age and gender. To do so, a classic speech-in-noise perception test involving naïve participants was conducted in three modalities: AV (auditory-visual), A (auditory only) and V (visual only). The results suggest that visual information is relatively preserved in speech produced by people with DS despite their anatomical and motor specificities. Moreover, it improves overall intelligibility. This however depends on phonetic feature and consonant. The results are discussed in relation to the main questions raised in the introduction and considering previous works and methodological aspects.

Is visual information preserved in speech produced by people with DS? Does it improve auditory intelligibility?

Previous work extensively reported that speech intelligibility is almost always, even though to various degrees, impaired in people with DS, especially when the acoustic signal is perceived alone (e.g. Bunton et al., 2007; Kent & Vorperian, 2013; Kumin, 2006). This is confirmed by our own findings. Our first aim was to evaluate the quality of the visual information in speech

produced by people with DS. In this study, perception scores in V are equivalent for speakers with DS and typical speakers. This could be counter-intuitive considering craniofacial, muscle and vocal-tract anomalies in DS (e.g. Kent & Vorperian, 2013; Latash et al., 2008; Macho et al., 2014). It could however be accounted for both by the inter-speaker variability in the "quality" of the visual information usually observed in typical speakers (e.g. all speakers do not provide clear visual features, Mallick, Magnotti, & Beauchamp, 2015) and the listeners' ability to make use of this information (e.g. all listeners are not good at lip reading: Bernstein, Demorest, & Tucker, 2000; Mallick, Magnotti, & Beauchamp, 2015). Note that even if they were low, the intelligibility scores in the V modality were still above chance confirming that this modality does carry information per-se. We then wanted to evaluate whether, in a situation in which processing the visual information is crucial to perceive speech (speech-in-noise), this information could improve the intelligibility of speakers with DS. Our analyses suggest that this is the case: VCVs produced by people with DS are globally more accurately perceived in AV than in A. Just as for typical speakers, it thus appears that seeing the speaker's face is beneficial to identify VCVs uttered by speakers with DS. Note that visual enhancement is similar in both groups showing that perceivers benefit as much of the visual information for perceiving DS and control speakers. Visual enhancement and its comparison between groups could be further investigated using a participant and speaker specific adaptive signal to noise ratio procedure (as in Bernstein et al., 2004; Sommers et al., 2005) in order to equate performance levels in A (for example at 50%).

All together, the latter observations suggest that the visual speech information is relatively preserved in speakers with DS, despite the anatomical and motor specificities caused by DS, and that it can be beneficially used to better perceive DS speech. Speech rehabilitation could use such findings and involve the visual modality to a greater extent both in speech evaluation and

rehabilitation protocols. This idea is similar to that suggested in Hustad et al. (2007) for people with dysarthria. Speech therapists could more systematically train speakers with DS to enhance their visual speech cues by using, for example, systematic visual feedback: simple video or ultrasound biofeedback already used with children with different types of speech disorders (Cleland, Scobbie, & Wrench, 2015). Further work should however be conducted in order to extend our results to more natural speech material such as words and sentences. It would also be interesting to compare the contribution of visual information to perception by unfamiliar listeners, such as in this study, to perception by familiar listeners such as parents, teachers, and/or professional staff. This would indeed make it possible to evaluate whether the people familiar with speakers with DS spontaneously use the visual information to improve their understanding of the person or whether it would be worthwhile to train them to do so (as for example can be successfully done with hearing impaired patients, Massaro & Light, 2004). It would also be interesting to run speaker-specific studies to assess to which extent anatomical and neuromuscular specificities in DS speakers influence the visual correlates of their speech as well as how these specificities interact with acoustic properties. Note that these effects could not be reliably assessed in the current study due to the small number of repetitions for each speaker imposed by experimental timing constraints. An exploratory by-speaker analysis of our dataset suggests that intelligibility scores in A are similar between the speakers with DS and always smaller than for the typical speakers. By contrast, the speakers from both groups were less distinguishable (control vs. DS) in AV and even less in V.

If visual information improves the perception of VCV sequences produced by speakers with DS, confusion and information transmission analyses clearly show that the contribution of vision depends on the consonant and especially on its phonetic features.

Which phonetic features are specifically impaired in consonants produced by speakers with DS? Is this effect modality dependent and how?

The confusion matrix obtained for the control speakers in the A modality was compared to the historical consonant confusion matrix published by Miller and Nicely (1955, M&N) for SNR = -6dB as well as to that published by Phatak, Lovitt and Allen (2008, PLA) for SNR = -6dB. This resulted in less than 7% mean differences between their confusion matrices and ours (M&N: mean = 6.8%, standard deviation = 10.1% – PLA: 6.4%, 11.2%). Our results are thus consistent with previous results especially considering that the language is different (English vs. French) as well as the noise type (white vs. cocktail-party noise) and that the two studies used CV sequences (vs. VCV in the present study).

In typical speakers, labial/bi-labial consonants were usually better identified in AV than in A and V (AV>A~V), while coronals rather followed the trends AV~A>V or AV~A~V and plosive dorsals the trend AV>A>V. The visual information is indeed greater for labials than for coronals. Surprisingly the perception of /k/ and /g/ however appears to benefit from vision. Voicing was better transmitted in A than in V as classically observed (Alm et al., 2009; Summerfield, 1987). Manner of articulation followed a similar trend, even though differences were less dramatic than for voicing. Similar observations were made for speakers with DS for both place and manner of articulation. The main inter-group difference is observed for voicing: whereas it is as poorly transmitted for both groups in V, it is dramatically less well transmitted for DS than control speakers in A and AV. It therefore appears that speakers with DS have issues in producing voicing. This result is important since there are not a lot of studies reporting intelligibility of voicing in DS. Borghi (1990) had already signaled voicing errors in speech produced by people with DS (see also Bunton et al. (2007) even though results are imprecise concerning that matter).

Smith and Stoel-Gammon (1983) also put forward devoicing of final stops in 5 children with DS. Kent and Vorperian (2013) report "increased noise in phonation" for DS speech. Also note that whichever modality and speaker group, unvoiced responses were provided more frequently than voiced ones. This is contrary to previous findings showing that unvoiced consonants seem to be less robust to noise for typical speakers, especially babble noise (Alm et al., 2009). Interestingly however, the latter tendency was stronger for DS than typical speakers especially in AV and A. This once again puts forward the fact that voicing would be particularly affected in the speech produced by people with DS and more specifically that they would have a tendency to devoice voiced consonants. Since voicing is poorly transmitted in V, as for typical speakers (e.g. Binnie, Montgomery, & Jackson, 1974), adding vision cannot compensate for it. Note however that some researchers suggest that voicing information can partially be recovered through the visual modality even though the larynx is not directly visible (Files, Tjan, Jiang, & Bernstein, 2015).

Some effects appear to be consonant-specific, suggesting interactions between phonetic feature and speaker group. This is particularly the case for /d/, poorly identified for speakers with DS in all the conditions and mainly mistaken for /t/. This was also the case for typical speakers but only in V, which is trivial since voicing is not well transmitted in V. A possible explanation for this could be an effect of relative frequency of the consonants (C) in an aCa context in French words: if the /ata/ sequence is relatively more frequent in French words than the /ada/ sequence, listeners may expect more /t/ than /d/ resulting in a bias in responding /t/ rather than /d/. Note however that the effect is speaker group and modality dependent (e.g. for DS speakers, /d/ responses were rare in AV while frequent in V) which invalidates the argument. Moreover, we found no significant correlation between the frequency of consonant occurrence in French in an aCa context (Freq_lang) and the frequency of occurrence of these consonants in the participants' responses

(Freq_resp; all conditions together, R=.11). Freq_lang was computed as the frequency of occurences of each aCa in all French words regardless of the position in the word relative to the sum of the frequencies of occurences of all aCa (movie+book frequency, Lexique database, New, Pallier, Ferrand, & Matos, 2001). Freq_resp was computed as the number of each consonant answers divided by the total number of responses. The finding that voicing would be particularly impaired in DS could have implications for speech therapy protocols. Andrade et al. (2014) review and compare several techniques used to train voicing, some of which could easily be used with people with DS, such as the straw exercise.

It was also observed that nasal responses were less frequent than responses corresponding to other manners of articulation for both speaker groups. It may be the case that nasality was particularly affected by the type of noise used.

Finally, an intriguing finding is that, whereas place information is transmitted as efficiently in both speaker groups in V, it drops largely in A but only for speakers with DS. It could be hypothesized that due to problems of relative macroglossia and difficulties in tongue motor control, speakers with DS try to articulatorily compensate using their lips. This would compensate for place information transmission in V resulting in no difference with control speakers. This would however not operate anymore in A resulting in poorer place information transmission than for control speakers. Similar observations, but in the reverse direction (compensation using the tongue), have already been observed for blind speakers (Ménard, Trudeau-Fisette, Côté, & Turgeon, 2016).

Potential influence of methodological limitations

One could put forward several methodological issues in the present study that could influence its results. In particular, repetition tasks were used both to record the stimuli and to collect participant responses in the perceptual test.

Involving speakers with an intellectual deficiency required specific adaptations of experimental procedures. In previous work evaluating speech production in speakers with DS using a reading task and real words, repetition was required in some trials when speakers failed to read correctly (cf. Bunton, Leddy, & Miller, 2007, p. 4: "If a word was mispronounced, the speaker was asked to repeat the word, if a second error occurred, the investigator read the word aloud and asked the speaker to repeat it."). Based on this report, we considered that repetition for all speakers was a good compromise to avoid bias between trials and speaker groups, and to design a more inclusive study. However, this task may have resulted in wrong phonetic identification of the stimulus rather than pronunciation errors: did the speakers, in particular those with DS, produce the expected stimulus? Could the lower scores observed in the perceptual test for DS be explained by wrong identification of the target utterance to be produced rather than articulation issues in achieving the target? This possible bias was first addressed in the recording procedure. As described in the methods section, each VCV sequence repetition was prompted by three different audio stimuli produced by three different speakers. Each stimulus was also played several times when required until the two experimenters judged that the participant did her best to produce the correct VCV. This reduces potential misinterpretations of what to repeat. Then, only the best repetition (e.g. the closest to the target as judged by three of the authors) was selected for the perceptual test. If this methodological approach does not exclude the bias completely (i.e. it is still possible that both typical, and even more so DS speakers, wrongly identified the target VCV), it can't account for the main results of the study (i.e. AV>A for both speaker groups and

V in DS ~ V in typical). If speakers with DS produced the wrong sequence because they misinterpreted the prompt they heard, they would not be better identified in AV than in A, and identification in V would have been poorer for DS than typical speakers.

On the other hand, the repetition task used to collect participant responses during the perceptual test may have induced two problems: 1. the participant did not manage to re-produce what she had just heard; 2. this procedure requires post-coding of the responses involving interpretation by the coder. To address 1, we could have used a forced-choice task or written transcription. We did not want to choose the first option because we wanted to be sure of what the participants actually perceived (which may not be in the alternative choices, see Bunton, Leddy, & Miller, 2007, for similar issues with real words). The fact that we observed "other" responses (corresponding to none of the 16 target consonants) confirms that forced choice wouldn't have necessarily assessed true perception. The second option was also discarded to avoid spelling ambiguities. To address problem 2, we used multiple coding by three coders (as described in the methods section). Strong agreement between coders shows that response coding may have only had minor influence on the results.

Conclusion

The current work provides new insight relevant to the study of speech intelligibility in people with DS and to the development of speech therapy for these persons. First, it shows that despite anatomical and motor specificities, the visual speech information seems preserved in the speech of people with DS. Then, it appears that speech produced by speakers with DS can be better understood when seeing the speaker's face rather than just listening to her and this visual benefit is as important as for typical speakers. Part of the solution to the speech intelligibility deficit in people with DS could come from the listener herself. People indeed tend not to look straight at

their interlocutor with DS, often out of shyness or discomfort, but this behavior actually impairs their chances to understand what their interlocutor tries to tell them. Our results also show that the contribution of visual information to the perception of consonantal features is particularly true for place of articulation and to a lesser extent for manner of articulation. Voicing appears to be the most altered phonetic feature in DS with a tendency towards devoicing. Vision cannot, or at the best barely, contribute to compensate for this voicing deficit. Previous work evaluating phonetic intelligibility in adult speakers with DS mostly studied word identification using minimal pair multi-choice tests or transcriptions, conducted in the auditory modality only and involving native speakers of English. The present study involved native speakers of French and suggests that visual information should be considered when evaluating speech intelligibility especially in speakers with DS. Further studies involving more naturalistic speech material, investigations of speaker-specific effects and noise effects are now required to better understand the potential role of visual information in the perception of speakers with DS. Cross-linguistic studies may also help in identifying difficulties specifically related to DS.

Acknowledgement

This research has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013 Grant Agreement no.339152- “Speech Unit(e)s”) and from the FIRAH foundation (International Foundation of Applied Disability Research). It was approved by the CERNI ethics committee of Grenoble Alpes University (IRB00010290 COMUE Grenoble Alpes University IRB#1 – approval number: 2014-03-11-41) and by the ethical comity of the FIRAH. We thank the ARIST (Down Syndrome Research and Social Integration Association), the ESAT-SAJ (Institution and Service through Work - Day Activity Service) and the speakers who participated in this study and their families.

References

Alm, M., Behne, D. M., & Wang, Y. (2009). Audio-visual identification of place of articulation and voicing in white and babble noise. *The Journal of the Acoustical Society of America*, *126*(1), 377–387. https://doi.org/10.1121/1.3129508

Andrade, P. A., Wood, G., Ratcliffe, P., Epstein, R., Pijper, A., & Svec, J. G. (2014). Electroglottographic study of seven semi-occluded exercises: LaxVox, straw, lip-trill, tongue-trill, humming, hand-over-mouth, and tongue-trill combined with hand-over-mouth. *Journal of Voice*, *28*(5), 589–595. https://doi.org/10.1016/j.jvoice.2013.11.004

Arumugam, A., Raja, K., Venugopalan, M., Chandrasekaran, B., Kovanur Sampath, K., Muthusamy, H., & Shanmugam, N. (2015). Down syndrome - A narrative review with a focus on anatomical features. *Clinical Anatomy*, *29*(5), 568–577. https://doi.org/10.1002/ca.22672

Barnes, E., Roberts, J., Long, S. H., Martin, G. E., Berni, M. C., Mandulak, K. C., & Sideris, J. (2009). Phonological accuracy and intelligibility in connected speech of boys with fragile X syndrome or Down syndrome. *Journal of Speech, Language, and Hearing Research : JSLHR*, *52*(August), 1048–1061. https://doi.org/10.1044/1092-4388(2009/08-0001)

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. *Speech Communication*, *44*(1–4), 5–18. https://doi.org/10.1016/j.specom.2004.10.011

Bernstein, L. E., Demorest, M. E., & Tucker, P. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233–252. https://doi.org/10.3758/BF03205546

Binnie, C. a, Montgomery, a a, & Jackson, P. L. (1974). Auditory and visual contributions to the

perception of consonants. *Journal of Speech and Hearing Research*, *17*(4), 619–630. https://doi.org/10.3758/bf03211678

Bittles, A. H., Bower, C., Hussain, R., & Glasson, E. J. (2007). The four ages of Down syndrome. *European Journal of Public Health*, *17*(2), 221–225. https://doi.org/10.1093/eurpub/ckl103

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Borghi, R. W. (1990). Consonant Phoneme, and Distinctive Feature Error Patterns in Speech. In *Clinical Perspectives in the Management of Down Syndrome* (pp. 147–152). New York, NY: Springer US. https://doi.org/10.1007/978-1-4613-9644-4_12

Borrie, S. A. (2015). Visual speech information: A help or hindrance in perceptual processing of dysarthric speech. *The Journal of the Acoustical Society of America*, *137*(3), 1473–1480. https://doi.org/10.1121/1.4913770

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Bunton, K., & Leddy, M. (2011). An evaluation of articulatory working space area in vowel production of adults with Down syndrome. *Clinical Linguistics & Phonetics*, *25*(4), 321–334. https://doi.org/10.3109/02699206.2010.535647

Bunton, K., Leddy, M., & Miller, J. (2007). Phonetic intelligibility testing in adults with Down syndrome. *Down Syndrome Research and Practice*, *12*(1), 1–4. https://doi.org/10.3104/editorials.2034

Campbell, R. (2008). The Processing of Audio-Visual Speech: Empirical and Neural Bases. *Source: Philosophical Transactions: Biological Sciences Phil. Trans. R. Soc. B*, *363*(363),

1001–1010. https://doi.org/10.1098/rstb.2007.2155

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *27*(1), 115–127. https://doi.org/10.1037//0096-1523.27.1.115

Chapman, R., & Hesketh, L. (2001). Language, cognition, and short-term memory in individuals with Down syndrome. *Down Syndrome Research and Practice*, *7*(1), 1–7. https://doi.org/10.3104/reviews.108

Cleland, J., Scobbie, J. M., & Wrench, A. A. (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics & Phonetics*, *29*(8–10), 575–597. https://doi.org/10.3109/02699206.2015.1016188

Cleland, J., Wood, S., Hardcastle, W., Wishart, J., & Timmins, C. (2010). Relationship between speech, oromotor, language and cognitive abilities in children with Down's syndrome. *International Journal of Language & Communication Disorders*, *45*(1), 83–95. https://doi.org/10.3109/13682820902745453

Connaghan, K. P., & Moore, C. A. (2013). Indirect Estimates of Jaw Muscle Tension in Children With Suspected Hypertonia, Children With Suspected Hypotonia, and Matched Controls. *Journal of Speech, Language & Hearing Research*, *56*(1), 123–136. https://doi.org/10.1044/1092-4388(2012/11-0161)

Crosley, P. A., & Dowling, S. (1989). The relationship between cluster and liquid simplification and sentence length, age, and IQ in Down's syndrome children. *Journal of Communication Disorders*, *22*(3), 151–168. https://doi.org/10.1016/0021-9924(89)90013-0

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2003.08.014

Files, B. T., Tjan, B. S., Jiang, J., & Bernstein, L. E. (2015). Visual speech discrimination and identification of natural and synthetic consonant stimuli. *Frontiers in Psychology*, *6*, 878. https://doi.org/10.3389/fpsyg.2015.00878

Fournier, D. A., Skaug, H. J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M. N., … Sibert, J. (2012). AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software*, *27*(2), 233–249. https://doi.org/10.1080/10556788.2011.597854

Grant, K. W., & Seitz, P. F. P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America*, *108*(3), 1197–1208. https://doi.org/10.1121/1.422512

Grant, K. W., Tufts, J. B., & Greenberg, S. (2007). Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *The Journal of the Acoustical Society of America*, *121*(2), 1164–1176. https://doi.org/10.1121/1.2405859

Guimaraes, C. V. A., Donnelly, L. F., Shott, S. R., Amin, R. S., & Kalra, M. (2008). Relative rather than absolute macroglossia in patients with Down syndrome: implications for treatment of obstructive sleep apnea. *Pediatric Radiology*, *38*(10), 1062–7. https://doi.org/10.1007/s00247-008-0941-7

Haute Autorité de Santé. (2015). *Les performances des tests de dépistage de la trisomie 21 foetale par analyse de l'ADN libre circulant*.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous Inference in General Parametric Models. *Biometrical Journal*, *50*(3), 346–363. https://doi.org/10.1002/bimj.200810425

Hustad, K. C., & Cahill, M. A. (2003). Effects of presentation mode and repeated familiarization on intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, *12*(2), 198–208. https://doi.org/10.1044/1058-0360(2003/066)

Hustad, K. C., Dardis, C. M., & McCourt, K. A. (2007). Effects of visual information on intelligibility of open and closed class words in predictable sentences produced by speakers with dysarthria. *Clinical Linguistics & Phonetics*, *21*(5), 353–67. https://doi.org/10.1080/02699200701259150

Katz, G., & Lazcano-Ponce, E. (2008). Intellectual disability: definition, etiological factors, classification, diagnosis, treatment and prognosis. *Salud Pública de México*, *50*(2), s132–s141. https://doi.org/10.1590/S0036-36342008000800005

Keintz, C. K., Bunton, K., & Hoit, J. D. (2007). Influence of visual information on the intelligibility of dysarthric speech. *American Journal of Speech-Language Pathology*, *16*(3), 222–234. https://doi.org/10.1044/1058-0360(2007/027)

Kent, R. D., & Vorperian, H. K. (2013). Speech Impairment in Down Syndrome: A Review. *Journal of Speech, Language & Hearing Research*, *56*(1), 178–210. https://doi.org/10.1044/1092-4388(2012/12-0148)

Kumin, L. (2006). Speech intelligibility and childhood verbal apraxia in children with Down syndrome. *Down's Syndrome, Research and Practice*, *10*(1), 10–22. https://doi.org/10.3104/reports.301

Kumin, L. (2012). *Early communication skills for children with Down Syndrome: A guide for*

parents and professionals. Woodbine House.

Latash, M., Wood, L., & Ulrich, D. (2008). What is currently known about hypotonia, motor skill development, and physical activity in Down syndrome. *Down Syndrome Research and Practice (Online)*. https://doi.org/doi:10.3104/reviews.2074

Loane, M., Morris, J. K., Addor, M.-C., Arriola, L., Budd, J., Doray, B., … Dolk, H. (2013). Twenty-year trends in the prevalence of Down syndrome and other trisomies in Europe: impact of maternal age and prenatal screening. *European Journal of Human Genetics*, *21*(1), 27–33. https://doi.org/10.1038/ejhg.2012.94

Macho, V., Andrade, D., Areias, C., Coelho, A., & Melo, P. (2014). Comparative Study of the Prevalence of Occlusal Anomalies in Down Syndrome Children and Their Siblings. *British Journal of Medicine and Medical Research*, *4*(35), 5604–5611. https://doi.org/10.9734/BJMMR/2014/12688

Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin & Review*, *22*(5), 1299–1307. https://doi.org/10.3758/s13423-015-0817-4

Massaro, D. W. (1987). Speech perception by ear and eye. In *Hearing by eye: The psychology of lip-reading* (pp. 53–58). https://doi.org/doi:10.4324/9781315799742

Massaro, D. W., & Light, J. (2004). Using visible speech to train perception and production of speech for individuals with hearing loss. *Journal of Speech, Language, and Hearing Research*, *47*(2), 304–320. https://doi.org/10.1044/1092-4388(2004/025)

Ménard, L., Trudeau-Fisette, P., Côté, D., & Turgeon, C. (2016). Speaking clearly for the blind: Acoustic and articulatory correlates of speaking conditions in sighted and congenitally blind

speakers. *PLoS ONE*, *11*(9). https://doi.org/10.1371/journal.pone.0160088

Meyer, C., Theodoros, D., & Hickson, L. (2016). Management of swallowing and communication difficulties in Down syndrome: A survey of speech-language pathologists. *International Journal of Speech-Language Pathology*, *19*(1), 1–12. https://doi.org/https://doi.org/10.1080/17549507.2016.1221454

Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonant. *The Journal of the Acoustical Society of America*, *27*(2), 338–352. https://doi.org/10.1121/1.1907526

Moura, C. P., Cunha, L. M., Vilarinho, H., Cunha, M. J., Freitas, D., Palha, M., … Pais-Clemente, M. (2008). Voice parameters in children with Down syndrome. *Journal of Voice*, *22*(1), 34–42. https://doi.org/10.1016/j.jvoice.2006.08.011

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet□: LEXIQUE ™ // A lexical database for contemporary french□: LEXIQUE ™. *L'année Psychologique*, *101*(3), 447–462. https://doi.org/10.3406/psy.2001.1341

Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., … Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004-2006. *Birth Defects Research Part A - Clinical and Molecular Teratology*, *88*(12), 1008–1016. https://doi.org/10.1002/bdra.20735

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169–181. https://doi.org/10.1016/j.cortex.2015.03.006

Phatak, S. a, Lovitt, A., & Allen, J. B. (2008). Consonant confusions in white noise. *The Journal*

of the Acoustical Society of America, *124*(2), 1220–1233. https://doi.org/10.1121/1.2913251

R Development Core Team. (2008). R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In *Hearing by Eye: The Psychology of Lip-reading* (pp. 97–114).

Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. *The Journal of the Acoustical Society of America*, *103*(6), 3677–89. https://doi.org/10.1121/1.423069

Rosin, M. M., Swift, E., Bless, D., & Vetter, D. K. (1988). Communication profiles of adolescents with Down syndrome. *Journal of Childhood Communication Disorders*, *12*(1), 49–64. https://doi.org/10.1177/152574018801200105

Ross, L. A., Saint-amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do You See What I Am Saying□? Exploring Visual Enhancement of Speech Comprehension in Noisy Environments. *Cerebral Cortex*, *17*(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Rupela, V., Velleman, S. L., & Andrianopoulos, M. V. (2016). Motor speech skills in children with Down syndrome: A descriptive study. *International Journal of Speech-Language Pathology*, *18*(5), 483–492. https://doi.org/10.3109/17549507.2015.1112836

Schwartz, J. L., Berthommier, F., & Savariaux, C. (2004). Seeing to hear better: Evidence for early audio-visual interactions in speech identification. *Cognition*, *93*(2), B69–B78. https://doi.org/10.1016/j.cognition.2004.01.006

Smith, B. L., & Stoel-Gammon, C. (1983). A longitudinal study of the development of stop consonant production in normal and Down's syndrome children. *The Journal of Speech and Hearing Disorders*, *48*(2), 114–118.

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*(1), 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Sommers, M. S., Tye-murray, N., & Spehar, B. (2005). Auditory-Visual Speech Perception and Auditory- Visual Enhancement in Normal-Hearing Younger and Older Adults. *Ear and Hearing*, *26*(3), 263–275.

Sommers, R. K., Patterson, P., & Wildgen, P. L. (1988). Phonology of Down syndrome speakers, ages 13-22. *Journal of Childhood Communication Disorders*, *12*(1), 65–91. https://doi.org/10.1177/152574018801200106

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215. https://doi.org/10.1121/1.1907309

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In *Hearing by Eye: The Psychology of Lip-reading* (pp. 3–51). https://doi.org/10.2307/1423237

Timmins, C., Cleland, J., Wood, S. E., Hardcastle, W. J., & Wishart, J. G. (2009). A perceptual and electropalatographic study of /ʃ/ in young people with Down's syndrome. *Clinical Linguistics & Phonetics*, *23*(12), 911–925. https://doi.org/10.3109/02699200903141271

Timmins, C., Hardcastle, W. J., Wood, S., & Cleland, J. (2011). An EPG analysis of /t/ in young

people with Down's syndrome. *Clinical Linguistics & Phonetics*, *25*(11–12), 1022–1027. https://doi.org/10.3109/02699206.2011.616981

Toğram, B. (2015). How do families of children with down syndrome perceive speech intelligibility in Turkey? *BioMed Research International*, *2015*, 11 pages. https://doi.org/10.1155/2015/707134

Xue, S. A., Kaine, L., & Ng, M. L. (2010). Quantification of vocal tract configuration of older children with Down syndrome: A pilot study. *International Journal of Pediatric Otorhinolaryngology*, *74*(4), 378–383. https://doi.org/10.1016/j.ijporl.2010.01.007

Zeiliger, J., Serignat, J., Autessere, D., & Meunier, C. (1994). Bd_bruit, une base de données de parole de locuteurs soumis à du bruit. *Actes Des Xèmes JEP*, 287–290.

Tables and Figures

Table 1. Characteristics of the speakers with DS and their control counterparts: identifier-age-gender.

Table 2. Phonetic features of the 16 consonants: Voicing: unvoiced (0), voiced (1); Place of articulation: labial (L), coronal (C), dorsal (D); Manner of articulation: plosive (P), fricative (F), nasal (N) and other (O).

Table 3. Examples of response transcriptions for stimulus /ada/ and associated accuracy scores for the entire VCV (Correct VCV) or the consonant-only (Correct C), see text for details.

Table 4. Distribution of responses for each speaker group (Ctr vs. DS) as a function of their type (Resp. type): correct responses (Correct) and responses with an error on: the consonant only (Err. Cons.), the consonant and another item (Err. Cons. + Other), another item only (Err. Other). Conf. Err. is the number

of errors involving a confusion between consonants. Percentages relative to the whole dataset (all responses) are provided in parenthesis.

Table 5. Details of the coefficient and variance estimates of the model used in Analysis 1 (*Prob_correct_VCV ~ Modality * Speaker_group * Stimulus + Modality * Speaker_group | Participant + Modality * Speaker_group | Speaker*).

Table 6. Results (estimate, standard error, z-value and p-value) of multiple comparisons testing between speaker group differences as a function of modality and between modality differences as a function of speaker group. These were obtained from the logistic regression model corresponding to Analysis 1 (*Prob_correct_VCV ~ Modality * Speaker_group * Stimulus + Modality * Speaker_group | Participant + Modality * Speaker_group | Speaker*). Stars highlight significant differences (*: p<.05, **: p<.01, ***: p<.001).

Figure 1. Organization of a trial and sequencing of trials within a block (see text for details). Colored screens and video stimuli were of the same size, the figure zooms on the stimulus screen for space reason. Color names of intermediate screens are written in brackets for gray scale printing.

Figure 2. Probability of correct VCV responses (*Prob_correct_VCV*) averaged across participants as a function of *Modality* and *Speaker_group*. Error bars are between-subject 95% confidence intervals. Stars and connecting lines show significant differences (*: p<.05, **: p<.01, ***: p<.001).

Figure 3. AV gain (mean across participants) relative to performance in the A only modality as a function of speaker group. Error bars are between-subject 95% confidence intervals.

Figure 4. Confusion matrices for each modality and speaker group. Each cell, $m_{i,j}$, corresponds to the number of times response j was provided for stimulus i, all participants and speakers together. The number in bold on each line corresponds to the most frequent response for a given consonant. Color codes indicate the error-type for the different features (see the legend below the figure). 'Other' responses

correspond to cases in which the response was not one of the 16 consonants (e.g. cluster, no consonant identified, no response provided, ambiguous response…).

Figure 5. Percentage of transmitted information averaged across participants as a function of modality and speaker group. Significant effects between speaker groups are shown by connecting lines and stars (*: $p<.05$, **: $p<.01$, ***: $p<.001$).

Figure S3-1. Probability of correct VCV responses (*Prob_correct_VCV*) averaged across participants for each *Modality* as a function of *Speaker_group* and presentation order (*Pres_order*). Error bars are between-subject 95% confidence intervals. Horizontal lines and stars highlight significant differences between conditions ($p<.05$).

Supplementary material

Supplementary material S1. General information about intelligibility and orofacial specificities of the four speakers with DS.

[Insert SupMat 1]

Supplementary material S2. Details of transmitted information computation.

[Insert SupMat 2]

Supplementary material S3. Effect of presentation order on *Prob_correct_VCV*.

[Insert SupMat 3]

Supplementary material S4. Analysis of the probability of correct consonant identification as a function of experimental condition.

[Insert SupMat 4]